A Component Based Approach to Automatic Pronoun Resolution over Varying Writing Styles

By:
Carl Youngblood

Abstract -

Electronic media have made more information available than ever before. The result is that finding specific information can take longer than it should. Algorithms that can understand natural language could process an immense amount of written data into an organized system in a fraction of the time a person would take.

The work reported here focuses on one specific aspect of the natural language processing task: the resolution of pronoun references. The goal of this project was:

- To break apart the pronominal resolution process into a set of modules, each of which corresponds to one kind of information, extractable from the text containing the pronouns, that might help resolve which of the possible coreferents is correct.
- To evaluate the effectiveness of those individual modules when working with different text genres and writing styles.

The input to the pronoun resolution system consists of S-Trees built by Dan Bikel's statistical parser. The pronoun resolution system is structured as a series of modules. Each module encodes a type of information that may or may not help determine the referent of each pronoun being considered. The modules can be activated or deactivated without altering the overall program functionality. The program was designed this way to isolate each module's contribution to the pronoun referent resolution system. This allowed us to observe the contribution of the different information sources each module represents over a variety of text.

We ran a series of experiments where texts with different properties were analyzed by the pronoun-resolution system. In each test we measured the effectiveness of the entire system as well as individual modules. We also attempted to determine which modules played the most significant roles in all the cases where the correct pronoun referent was found. The data obtained from these experiments show that the module(s) with the greatest contribution to pronoun resolution accuracy depended on the style of text under analysis. Further, the data suggest that the same module will have a large contribution to pronoun resolution accuracy on all texts with similar writing styles. Sometimes the implemented modules were unable to accurately resolve a majority of the pronouns in a text. The data do not answer the question, “What is required to accurately find the referents for the pronouns in this text?” It is possible that some texts are sufficiently complex that no single module will be able to resolve the majority of its pronouns. However it is also possible that more sophisticated syntactic and structural constraints or semantic augmentation could result in a module that would correctly find the majority of the correct coreferents in a text. The result is an experimental framework that organizes future work into finding more pairings between writing styles and the information sources that work best with them.

One methodological issue is that a pronoun resolution system does not operate independently. It must interact with other standard natural language

system components such as a part of speech tagger, a parser, and a lexicon. If one does not want to build every piece, it is possible to use off the shelf components. However this leads to a host of system integration issues since there is no agreement on a canonical set of part of speech tags, or on how a parse tree for a sentence should be structured.

1. Introduction

Natural language processing is one of the most difficult problems facing computer science, yet the potential benefit if the problem is solved is huge. Although people are already capable of effectively perfect natural language processing, more information is produced every day than any individual can realistically process. Encoding this overwhelming amount of written text produced every day into machine readable form will allow us to access a much larger portion of it effectively by using computers to search through it rather than doing so ourselves.

1.1 What is Anaphora Resolution?

Coherent texts and dialogues contain multiple references to the same or related objects. Languages provide mechanisms to do this succinctly. Linguistic objects whose meaning can be determined only by linking to some other object are called anaphora (singular: anaphor). Some common anaphoric structures in English are: pronouns, one-anaphora, and part-whole relations.

Anaphora resolution refers to the NLP problem of identifying objects in a given text that are coreferent with the anaphor under consideration. In other words, finding which object an anaphor is referring to. Pronominal anaphora resolution is a subset of anaphora resolution and refers to the task of finding coreferents for pronouns.

1.2 Nature of the Problem

Consider the sentence “John took the apple and ate it.” The pronominal anaphor ‘it’ refers to the coreferent ‘apple’. If a natural language engine is going to build a meaning representation of this sentence it must have a way to link “it” to “apple”. There are several approaches to the problem of finding such coreferents. These approaches differ in the kinds of knowledge, both linguistic and domain specific, that they use. Based on what each approach uses it is categorized as either syntactic or semantic.

Finding a coreferent can rely on several different sources of information. This information can be categorized by where it comes from. Consider these text fragments:

"John took the apple and ate it."

Knowing that apples are something to be eaten suggests that apple is the proper coreferent rather than John. This is an example of using outside knowledge rather than the information contained in the structure of the sentence to solve the problem. Knowledge of what is edible is an example of a semantic information source. Note that this is not necessarily the easiest implementation to solve this problem. In this case a gender agreement filter would remove "John" from consideration since the pronoun "it" requires a neuter coreferent.

"The ball fell off the table. It rolled out the door."

Vs.

"The ball fell off the table. It was tilted toward the door."

In these examples "it" has two possible candidates "ball" and "table", neither of which are easily eliminated. Since the sentence structure of the two examples is almost exactly the same it would be very difficult to design a purely syntactic solution that can recognize the difference between the examples. This is a case where augmented semantic information is probably the best solution so that the pronoun resolution system can recognize that "ball" is something that would have "rolled" and "table" is something that would be "tilted".

"The boys saw the governator. They ran away from him."

Certain words contain inherent information. We know that "they" is plural and "him" is singular. From this we can deduce that "they" must refer to the plural coreferent "boys" and "him" must refer to the singular coreferent "governator". This is an example of the number agreement information source.

"I sat under the tree and watched the fireworks display. It was really amazing."

Because of the way English is typically structured, pronouns frequently refer to the coreferent that directly precedes them spatially. This is known as recency.

"The boy and the girl saw the bike in the window. He wanted it, but she did not."

We know that "He" and "she" must refer to people. There are two possible coreferents for "He" and "she" that are people, "boy" and "girl". If our pronoun resolution system also knows that "He" must refer to a male and "she" must refer to a female then the system can assign the masculine "He" to "boy" and the feminine "she" to "girl". This is an example of the gender agreement information source.

A second problem in this sentence is what "it" refers to. The two possible coreferents are originally "bike" and "window". One solution is to use semantic information to recognize that "window" is not realistically an object of desire for "boy". The syntactic solution is to model the focus of the sentence based on the structure. This allows the system to see that "bike" is the more likely coreferent for "it" because "window" is in a modifying prepositional phrase.

"My cat looked out the window and saw a moth. It batted at it."

This is an example where a semantic information source is very useful. Without knowing that "cat" is an object that would be have "batted" and that "moth" would be something "cat" would bat at, it would be very difficult to correctly assign what the first and second "it" refer to.

Different sources of information range in value. Sometimes they add no new information to whether or not a candidate is the correct answer, and sometimes they can always select the correct candidate without any assistance from other information sources. To test the importance of each source of information, they have been organized into isolated modules to compare their individual contributions. Each module is referred to as a constraint source because it imposes constraints on the pool of candidates being considered as possible coreferents. For example a constraint source can be passed a candidate that it can then throw out completely, suggest it is the correct answer, or suggest that it's the wrong answer. Suggesting means raising or lowering the "confidence score”. This score is just an aggregate value to represent the likelihood that the specific candidate is in fact the correct coreferent.

The type of text under analysis can affect what sort of information is available to the pronoun resolution system. Morphological information is whether a word is plural, capitalized, and has prefixes and/or suffixes. Typically this type of information is not abundant in English text compared to other languages. Taking advantage of this requires no actual knowledge of the meaning of the text so constraint sources based on morphological information are defined as syntactic.

A semantic constraint relies on knowledge of the actual meaning of the text under analysis. An example is the knowledge of whether or not something is alive. Using this information we can make a constraint source that will throw out candidates that are inanimate when the pronoun must refer to a living being, or vice versa.

The objective of this study is to compare the performance of these constraints as the text being analyzed by the pronoun resolution system varies. The current performance measure, overall coreferent resolution accuracy, doesn’t give a deep enough breakdown of the results. It is important to consider individual constraint source contributions to that overall accuracy as well. Otherwise it is impossible to discover the interaction between different text styles and constraint source performance.

2. Review of Literature

Contemporary research into anaphora resolution tends to take very specific approaches. General research on just pronominal anaphora resolution is rare because the problem is so difficult. Unfortunately the different approaches typically just report an overall accuracy measure to rate their success. There is rarely explanation of what was responsible for that accuracy.

Shalom Lappin and Herbert J. Leass published an approach that uses no semantics in *An Algorithm for Pronominal Anaphora Resolution*. Their approach is based on extracting information from the syntactic structure of the text and a model of attentional state. Their initial result is 86% accuracy overall when resolving third person pronouns and lexical anaphors. This breaks down into 74% accuracy for inter-sentence and 89% for intra-sentence cases. The testing material they used is drawn from the same corpus that their program was trained on.

Christopher Kennedy and Branimir Boguraev continued the work of Lappin and Leass and published *Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser.* They tried to improve the previous design to work when very little parse information is available, ideally just a part of speech tagger and a noun phrase extractor. This was largely accomplished by adding more constraints and fine tuning the preexisting ones. When their system was tested on twenty seven data sets drawn from sources such as magazine articles, press releases, and news stories, it performed with an overall accuracy of 75% (231 out of 304 third person anaphora resolved correctly).

Rebecca Watson, Judita Preiss, and Ted Briscoe present a perfect example of how anaphora resolution contributes to the larger task of NLP in *The Contribution of Domain-Independent Robust Pronominal Anaphora Resolution to Open-Domain Question-Answering.* As indicated the purpose of the research is to enrich available context in their stored text representations to improve the ability of their QA system to return results. They've tried to make their anaphora resolution domain independent by using only syntactic and structural information from the text. Their result was 73.2% accuracy.

Jose L. Vicedo and Antonio Ferrandez did similar work in *The Importance of Pronominal Anaphora resolution in Question Answering Systems*. The approach was syntactic using statistical knowledge, morphological agreement, and heuristics like preference for intra-sentence over inter-sentence coreferents. Their system achieved 87% accuracy.

These and other previous works all attempt to solve a subset of the overall problem. However their designs frequently use different approaches to the actual process, whether that means heuristics, syntactic information, or semantic information. There have been few tests to determine how much each constraint source is contributing to the overall program performance. Further, there is no cross comparison of algorithms used in the different designs.

Because different approaches have achieved good results on different types of text, there are two possibilities. One is that all the approaches are generally effective and just need to be augmented with more types of constraint sources and more tuning of the current constraint sources to achieve higher accuracy. The other possibility is that the different approaches happen to be designed to perform well on the specific test texts. If this is the case, it is important to research and catalog the importance of different anaphora resolution techniques based on the style of text it is being applied to. This can open up further work in creating designs that adaptively choose which information from constraint sources should be weighted more based on the text being processed.

3. Method -

First a children's story was chosen based on length, simplicity, and the presence of pronouns. The story was run through Bikel's parser unmodified, meaning the parser first POS tagged it and then parsed it. This creates a text output of S-Tree representations of each sentence to a text file. This is read as input to the pronoun resolution system, which rebuilds the text into S-Tree data structures. From here it is possible to choose which constraint sources to have active for each analysis run.

Later more tests were run on different selections of text. The purpose of using varying text samples was to compare the performance of individual constraint sources as the testing material changed. These data would allow us to make deductions about which constraint sources are most important when analyzing a certain style of text with the pronoun resolution system.

3.1 Program Structure

The main control flow is initialized by the basic wrapper class Exec. This initializes the input stream from the text file containing the S-Tree representations. Each time a new block of text is analyzed for anaphora resolution Exec creates a new instance of the Analysis class. The Analysis class contains an S-Tree data structure of the parsed text and the semantic constraints to be used in that analysis.

The reason we chose this design is that it leaves open the option for running more than one analysis text at a time. The Exec class could initialize multiple instances of the Analysis class which would then access the data structure classes to create specific trees for each instance to operate on. Then each Analysis class could run with different semantic constraints.

Program Structure

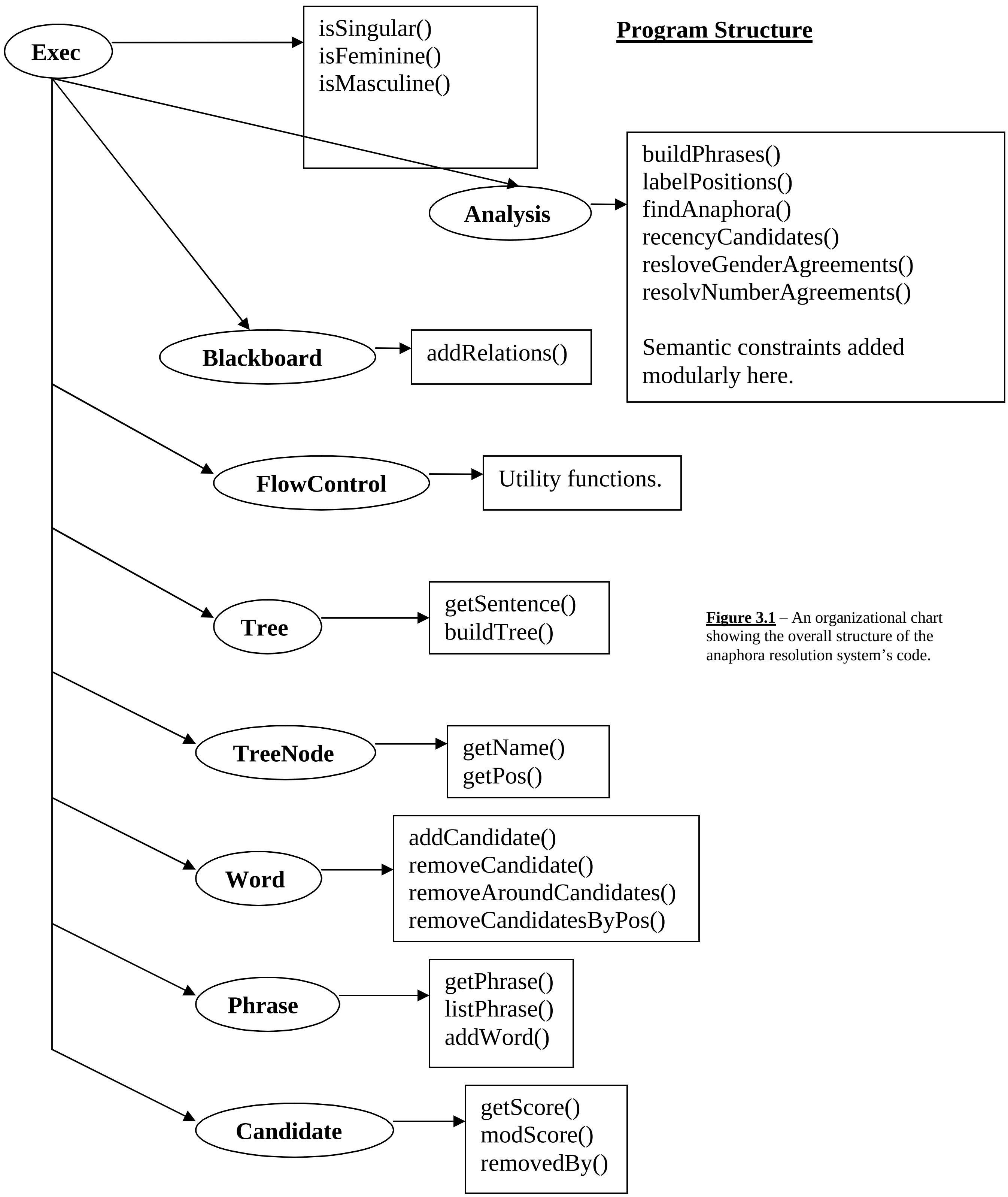

Figure 3.1 – An organizational chart showing the overall structure of the anaphora resolution system's code.

3.2 Blackboard

Pronoun resolution currently has no unified solution. As a result no single constraint source can resolve every pronoun's coreferent. To increase the accuracy of a system requires integrating multiple different constraint sources so that they can all operate on the same data. The most efficient way to do this is with a blackboard style of data organization.

A proposal-style module adds new candidates to the blackboard to be considered as the coreferent. Each candidate is proposed with a confidence score that is used later. These proposals are next filtered by constraints like gender and number agreement and the incompatible ones are removed from consideration. Finally candidates are evaluated by constraints like a discourse model or semantic type consistency. These modify the candidate's confidence score, and the candidate with the highest confidence score after being fully evaluated is returned as the actual coreferent.

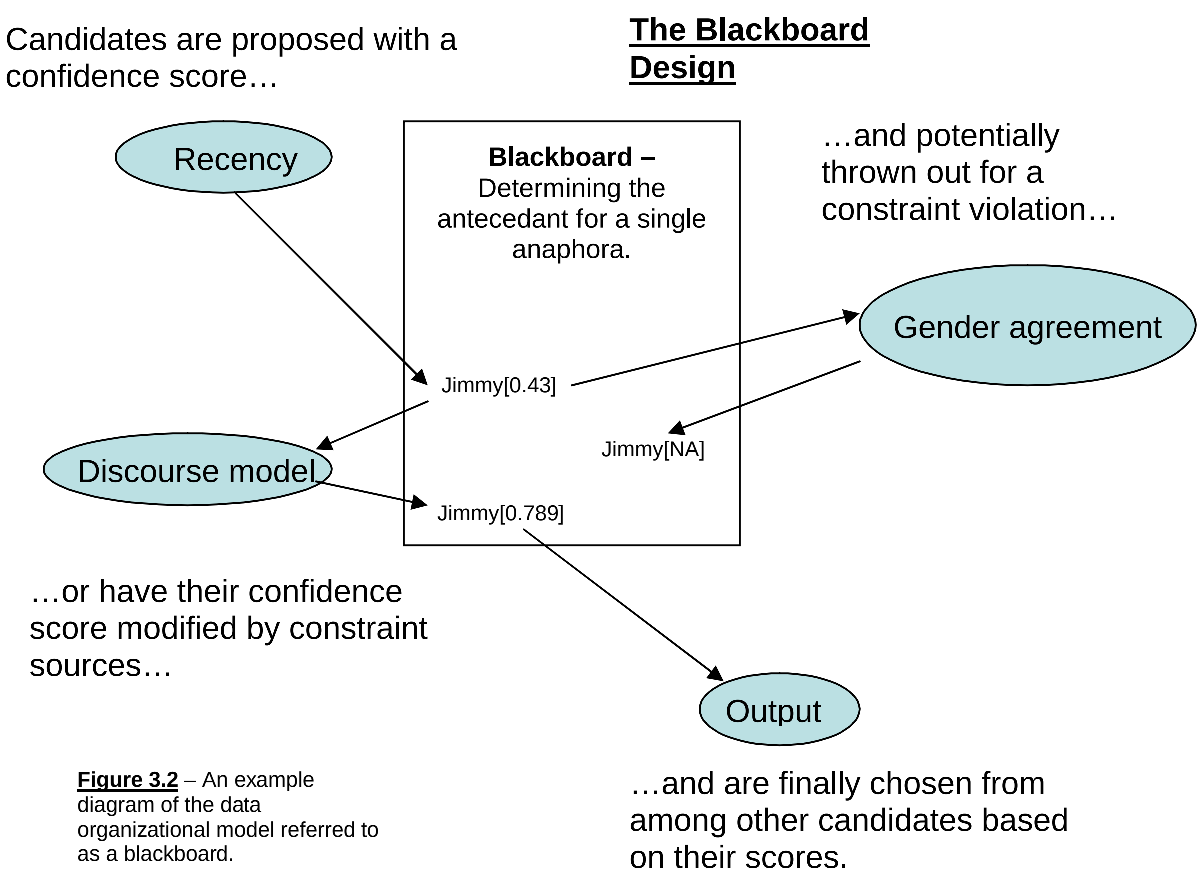

Figure 3.2 – An example diagram of the data organizational model referred to as a blackboard.

3.3 Constraint Sources

Originally only semantic constraint sources were going to be used. However syntactic constraint sources are included because they include some of the most effective techniques relative to their difficulty to implement. The modules available for use are:

Recency – A proposal source, recency moves backwards spatially through the text and adds noun phrases to the blackboard as candidates. The confidence score is set on proposal as a float value starting at one and exponentially decreasing to zero as the proposer reaches the beginning of the analyzed text.

Gender Agreement – GA compares the gender of candidate coreferents to the gender required by the pronoun being resolved. Any candidate that doesn't match the required gender of the pronoun is removed from further consideration.

Number Agreement – NA extracts the part of speech of candidates from the S-Tree. The part of speech label is checked for plurality. If the candidate is plural but the current pronoun being resolved doesn't indicate a plural coreferent the candidate is removed from consideration. The same process occurs for singular candidates which are removed if the pronoun being resolved requires a plural coreferent. This is an example of a constraint that relies on accurate part of speech tagging in the preprocessor.

Animistic Knowledge – Animistic knowledge filters candidates based on which ones represent living beings. Inanimate candidates are removed from consideration when the pronoun being resolved must refer to an animated coreferent, and animated candidates are removed from consideration for pronouns that must refer to inanimate coreferents.

4. Data and Discussion -

Some constraint sources rely on a database of words. An example is gender information in English, where nouns contain no inherent information about the gender of what they represent. The result is that gender agreement requires the pronoun resolution system to maintain a database that indicates whether each noun is masculine, feminine, or neuter. To simulate having a very large dictionary, words were added to the pronoun resolution system's database as they were encountered in test texts.

A standard test is based on comparing the contribution of different constraint sources to the overall accuracy of the correctly resolved pronouns. Previous work suggests that very few sources will do most of the work on any given text. This does not imply the sources doing the most work will always be the same. The modules that contribute the most will likely vary based on the style of writing in the text under analysis.

4.1 Experiment 1 – Children's Stories

Tested text:

Ten steps from the porch and twenty steps from the rose bushes growled Bluebeard in Jimmy's dream one night. There be treasure there! So the next day Jimmy began to dig. He dug until the hole was deep and the dirt pile was high. He kept digging. The hole got deeper and the dirt pile got higher. He dug until the hole was deepest and the dirt pile was at its highest. He sighed. I'm too tired. I can't dig anymore. Then he spied something but it was only one of Woofy's bones. Instead of treasure, all Jimmy had was a dog bone, a hole, and a big pile of dirt to fill it in with. He thought that pirate lied to me! But when Jimmy's mother saw what he had done, she clasped her hands and smiled a smile from here to Sunday. Oh, thank you, Jimmy. I always wanted a rhododendron bush planted just there. Here's $5.00 for digging that hole.

This first experiment used text from a children's story. Ideally this experiment represents a baseline performance since the story is a straightforward narrative style with extremely low sentence structure complexity.

Experiment 1 Constraint Sources Active	Correct Resolutions	Anaphora to Resolve	% Accuracy
Recency	4	14	28.57143
Recency, Number Agreement	4	14	28.57143
Recency, Number Agreement, Gender Agreement	10	14	71.42857
Recency, Number Agreement, Gender Agreement, Animistic Knowledge	13	14	92.85714

Table 4.1 – Results from experiment 1

The result was that NA had no contribution to accuracy, and GA increased the accuracy 42.9%. This text's pronouns are primarily third person gender indicative. This supports the idea of a few modules doing the most work, and suggests that a narrative style of text benefits most from gender resolution.

One error is from an issue where the parser gives the word "I'm" the incorrect part of speech as a noun. Fixing this would require improving the part of speech tagging during preprocessing.

Animistic knowledge increased the accuracy moderately. It benefited from the fact that the one main character of the story was frequently referred to so there were few competing candidates that were also animated to overwhelm the filter.

Clearly gender agreement was a valuable constraint source for this story. However it turns out to be even more valuable in later experiments despite the increased sophistication of the test text. This implies that simplicity is not required for GA to be successful. Instead the topics of the text seem more important, since this story was less about people and more about events than experiment three.

4.2 Experiment 2 – News Article

Tested text:

Scientists have long marveled over the dance of the bee. A little jitterbug seems to reveal to coworkers the location of a distant meal. But how and whether the dance really works has remained controversial. A new study confirms the dancing is a form of communication. Bees outfitted with tracking devices responded to the wiggling of one of their fellow foragers, who had just returned to the hive from some newfound bee vittles. The dance, which is performed on one of the honeycomb walls, is not an exact language, but it gets the job done. The central element of the choreography is a shimmy, or waggle, along a straight line. For emphasis, the bee repeats this move several times by circling around in a figure-8 pattern. The angle that the shimmy makes in relation to an imaginary vertical line is the direction to the food source with respect to the sun. For example, a waggle dance pointing towards 3 o'clock is bee talk for hey, there's food 90 degrees to the right of the Sun. This solar compass in honeybees was originally observed in the 1960s by the Nobel Prize winner Karl von Frisch. Later, it was noticed that the number of waggles in one figure-8 corresponds to the distance to the meal. These remarkable relations have been supported by other experiments, including one in which a mechanical bee danced for the hive and the real bees responded. But there have remained doubts as to whether the other bees could actually decipher the dancer's message. The dance isn't a trivial demonstration, but an abstract code, says J. R. Riley of Rothamsted Research, UK. One complication is that hives are dark and cramped, so other bees called recruits do not see the full pattern as human observers do. Furthermore, recruits tend to take longer to find the food than would be expected. Flying directly, it should only take them a minute or so, but they often don't find the feeder for 5 or 10 minutes, Riley told LiveScience. And sometimes they never find it. For this reason, some scientists have speculated that the waggle dance merely excites other bees, which then fly out of the hive searching for a scent trail left by the returning bee. Making a beeline to solve the controversy, Riley and colleagues strapped radar transponders to 19 dance spectators. The flight paths show that the bees make a beeline to the vicinity of the food source, but then fly around in a looping search pattern. Only two of the radar-tracked recruits actually found the food. Apparently, the dance gives incomplete instructions, and the bees rely on odors, colors, and other clues to hone in on the final location. Still, the dance gets them pretty close. On average, the recruits came within 18 feet of the food before switching to search mode. This was in spite of considerable wind drift which would have pushed them off course if they had not compensated, Riley said. To further investigate bee-havior, the team moved some recruits several hundred yards away from the hive and then released them. The displaced bees flew the prescribed direction and distance where they found nothing because their starting point was off. This is the most definitive proof that recruited bees read the waggle dance, since the transplanted bees chose the foretold trajectory without any of the possible other cues, smell, landscape, other bees that might exist along the true hive-to-feeder route.

The second experiment is on text from a LiveScience article. This article presents an entirely different challenge from the narrative story style. The writing style is much more complex and the pronouns lack the gender information present in experiment one and three.

Experiment 2

Constraint Sources Active	Correct Resolutions	Anaphora to Resolve	% Accuracy
Recency	0	13	0
Recency, Number Agreement	2	13	15.38462
Recency, Number Agreement, Gender Agreement	2	13	15.38462
Recency, Number Agreement, Gender Agreement, Animistic Knowledge	10	13	76.92308

Table 4.2 – Results from experiment 2

The program's performance indicates a few things. First, the complexity of the sentence structure in this article is much higher than a children's story, so recency drops from moderate to nonexistent accuracy. This appears to be from a much higher number of embedded clauses in sentences that introduce noun phrases that interfere by being proposed with a higher confidence score since they are spatially closer in the text to the pronoun being resolved.

Second, there is a noted lack of gender information. This is because the bees that are the topic of the article are always referred to with neuter pronouns. This represents a significant loss of information compared to the narrative story. The result is that gender agreement plays a very small role.

Animistic knowledge contributes significantly to overall accuracy. This makes sense in context of the article's writing style. A single topic, the behavior of bees, is being discussed through the entire article, and because bees are animated a large number of inanimate candidates were easily eliminated. This combines with the lack of competing animated candidates to result in dominance by this single constraint source.

Based on the general nature of the article, there are some modules that would increase the accuracy if added. First would be a heuristic to exclude candidates from embedded clauses immediately prior to the anaphora. Second would be a repetition heuristic that favors candidates which have been the coreferent for several pronouns already. Third is any sort of knowledge base or semantic constraint that can cover ideas like what exactly can fly or dance. The third considered addition could be considered less valuable then the first two though since the domain under which it contributes is restricted. However the proposed heuristics would also perform poorly if the writing style of the text being analyzed changed, so they are actually just as domain specific.

4.3 Experiment 3 – Romance Novel

Tested text:

Shocked by the older woman's frank appraisal, women her age weren't supposed to be looking at men, Enid exclaimed, Brown! A woman would have to be blind or dead not to appreciate him. Brown chuckled. I suppose that's why you want him to clothe himself, though. He's scarcely speaking to you, so I suppose you're not sharing his bed. Enid didn't need a confidante, nor did she need an advisor. She was perfectly capable of managing her life without help from anyone. Of course, she would have liked to tell someone MacLean's real problem, and see whether or not they thought he would ever forgive her, for it was she who had set off this frenzy of muscle-building. She had told him who he had been, and he hadn't liked hearing about his gambling, his cheating, and his wandering. He had been infuriated by her recitation of his crimes. And when she had said he'd abandoned her he had called her a fraud. She'd felt sorry for the man. He'd been so obviously flummoxed by her announcement. So she had let him abuse her and hadn't said a word, and what did she get in return? He could barely stand to gaze at her. They never held a real conversation any more.

The material from this experiment is a passage from the novel Lost in Your Arms by Christina Dodd. This text represents the same narrative format as experiment one's children's story. Unlike that story however, the sentence construction is much more complex. It contains the same prevalence of third person gender informative pronouns.

Experiment 3 Constraint Sources Active	Correct Resolutions	Anaphora to Resolve	% Accuracy
Recency	1	34	2.941176
Recency, Number Agreement	1	34	2.941176
Recency, Number Agreement, Gender Agreement	31	34	91.17647
Recency, Number Agreement, Gender Agreement, Animistic Knowledge	31	34	91.17647

Table 4.3 – Results from experiment 3

The major performance contributor here is very clear. Recency is only capable of correctly assigning a single coreferent while NA contributes nothing. However GA correctly chooses 30 coreferents out of 34, single handedly resolving over 88% of the pronouns. Recency's low contribution represents the increased complexity of sentence structure that was also present in the experiment two texts.

Gender agreement's huge role is explained by understanding that the most important things in a romance novel are the men and women. All the experimental passages are relatively the same length; however this one clearly has the most anaphora, and since they naturally tend toward referring to men and women the result is that gender agreement is an absolutely critical constraint source.

One of the anaphora that was incorrectly resolved was a pleonastic pronoun. This is a type of pronoun that doesn't have an antecedent, or the antecedent is implied or not to be referred to directly. A basic example is "It is clear the fighting will continue". The pronoun 'it' has no coreferent so it is classified as a pleonastic pronoun. Without a constraint source to handle this type of pronoun it was incorrectly assigned a candidate where there shouldn't have been one.

The other two errors involved a plural pronoun referring back to multiple proper noun coreferents. Example: "Jill and Jane ate breakfast. They chose cereal." In the current design the only proposal source is recency, and noun phrases are broken down into their components (if more than one) when they are proposed. Since the current setup only suggests a single correct coreferent for each pronoun, there is no way these can be resolved correctly.

Animistic knowledge doesn't contribute anything. Gender agreement already handles all the problems animistic knowledge would have solved. This is because in a way gender agreement is a subset of animistic knowledge. Since anything in this story that has a gender is also animated, all the inanimate candidates were already excluded by gender agreement.

The final results suggest that a standard narrative can be handled almost entirely by very few constraint sources despite structural complexity. The role of GA as the source of this success is important because it highlights how just adding more constraint sources is not necessarily as good as adding the right constraint sources.

4.4 Experiment 4 – Jane Austen

It was the beginning of February; and Anne, having been a month in Bath, was growing very eager for news from Uppercross and Lyme. She wanted to hear much more than Mary communicated. It was three weeks since she had heard at all. She only knew that Henrietta was at home again. Louisa, though considered to be recovering fast, was still at Lyme. She was thinking of them all very intently one evening, when a thicker letter than usual from Mary was delivered to her. To quicken the pleasure and surprise, with Admiral and Croft's compliments. The Crofts must be in Bath! A circumstance to interest her. They were people whom her heart turned to very naturally. Lady Russell saw either less or more than her young friend, for she saw nothing to excite distrust. She could not imagine a man more exactly what he ought to be than Elliot. She never did enjoy a sweeter feeling than the hope of seeing him receive the hand of her beloved Anne in Kellynch church, in the course of the following autumn.

Experiment four draws text from Jane Austen's Persuasion. The purpose of this experiment was to test the program on text that is still narrative but that doesn't follow normal grammar rules. It also presents the highest level of sentence structure complexity out of all the texts.

Experiment 4 Constraint Sources Active	Correct Resolutions	Anaphora to Resolve	% Accuracy
Recency	3	16	18.75
Recency, Number Agreement	4	16	25
Recency, Number Agreement, Gender Agreement	9	16	56.25
Recency, Number Agreement, Gender Agreement, Animistic Knowledge	9	16	56.25

Table 4.4 – Results from experiment 4

One result is that there is no clear major contributing constraint. Recency displays an abnormally high contribution of 18.75% considering the complexity of the text and number of embedded clauses. A possible explanation is that the coreferents that are the correct answer for many of the anaphora are repeated much more frequently than in other texts, giving recency more chances to provide a correct resolution.

Gender agreement is once again a solid contributor at 26.3%, although its performance is less than with other narrative styles. This is likely because of the sheer number of characters frequently referred to by their proper names. The result is that the gender agreement filter being overwhelmed by an excess of possible female coreferents.

Animism fails to contribute any new information. This is the same situation as the romance story experiment. Because of the bias towards including lots of gender information in the text the gender agreement filter covers everything animism would have caught.

Four of the errors (57% of the total errors) are due to an inability to keep track of multiple character names, this confuses the GA filter. One of the errors is a lack of method to handle pleonastic pronouns. The last two errors are a repeat of a plural anaphora having multiple character names as coreferents, see experiment three.

The results here run counter to the previous tests. Each constraint source contributes moderately to the accuracy but none of them dominates. In this case, achieving an acceptable performance would require the interaction of many constraint sources, all contributing relatively equal amounts. The test text represents unusually complicated structure and grammar which is difficult for both pronoun resolution and the preprocessing. Thus the need for a large number of constraint sources may fit because of the large variety of problems to be resolved.

4.5 Experiment 5 – Extended Children's Story

Tested text:

Mr. Coyote was getting very old and had to be more careful for his own safety. He had been walking for hours and hours through a beautiful valley when he came upon a large tree. Mr. Coyote was very tired and wanted to rest but he also needed to be safe. He kindly asked the tree, "Please open up so I can rest safely in your care". The tree opened up so that Mr. Coyote could go inside to rest, then it closed to keep him safe. Mr. Coyote slept for hours. When he woke up he could not remember what he had said to make the tree open. He said, "Let me out Mr. Tree", but nothing happened. He said, "Please let me out now!" and again nothing happened. The tree didn't even creak. Mr. Coyote knocked on the tree, but it would not open up. Mr. Tree was upset with Mr. Coyote for not having said *please* the first time he spoke to the tree! It let him rest a little longer. Because the birds heard Mr. Coyote banging on the inside of the tree, they came down to peck on the tree to help get him out. But they were too small and the tree was just too big! Finally Mr. Woodpecker came down and pecked a hole in the tree. Although it was a very small hole, it caused Mr. Woodpecker to get a bent beak! This meant he couldn't peck on the tree any more. Mr. Coyote put one hand out the hole but he could not fit through. He then tried his leg but still he could not fit through. He had to come up with a way to escape since Mr. Woodpecker's beak was now bent. Mr. Coyote knew there had to be a way. "Ah, come on you old ugly tree," he cried, "Just let me out!" But still nothing happened, just the silence around him. Mr. Coyote decided to take off his arms one at a time and put them through the small hole. He then put his legs through one at a time by taking them off. He put his body through by taking it off. This was working out fine. I'll show you Mr. Tree, you can't keep me in here, he thought. Next Mr. Coyote tried to put his head through the hole, but it was too big. His ears were in the way. So he took off his ears and put them through the hole. He again tried his head, but his eyes were too big. Mr. Coyote took his eyes off and put them through the hole. Mr. Raven saw the eyes and flew down to take them. Then Mr. Raven flew back up high in the tree with Mr. Coyote's eyes. They were such beautiful eyes, blue like the sky, and would be a treasure to put in his hiding spot! Mr. Coyote finally put his head through the hole. He then put himself back together. One piece at a time he became a whole coyote again. But after he put his head on he could not find his eyes. He was feeling all over. His ears were listening to hear him touch his eyes, but not a sound could be heard from his eyes. His fingers were being careful while feeling around, but still no eyes were found. Mr. Coyote knew he could not let the animals know he was blind. He felt his way to a wild rose bush; he then put two rose petals in for his eyes. This would cover the blindness for a little while, but he would have to keep looking for his eyes. Surely they were close by! Along came Mr. Snail who saw Mr. Coyote with the rose petals in his eyes. He asked Mr. Coyote, "Why do you have those rose petals in your eyes?" Mr. Coyote said, "Because they are very beautiful. They have lovely colors. You can try them if you want and I will hold your eyes." Mr. Snail took off his eyes. He put them into Mr. Coyote's hands and tried the rose petals in his eyes. Then Mr. Coyote put Mr. Snail's eyes into his head and ran off with his long tail wagging. To this day Mr. Snail is crawling with his head down looking for his eyes. And all coyotes have brown eyes instead of blue; this is because Mr. Coyote was naughty when he took Mr. Snail's eyes. And Mr. Raven still has those beautiful blue eyes in his secret hiding place, but he cannot return them because the secret hiding place was so secret not even Mr. Raven can find it!

Experiment five is run to reinforce the findings of experiment one and experiment three. It is much longer than the previous experimental passages with a many more anaphora. The sentence complexity is simple overall, however towards the last quarter of the text the complexity increases when Mr. Snail and Mr. Coyote begin a back and forth dialogue.

Experiment 5 Constraint Sources Active	Correct Resolutions	Anaphora to Resolve	% Accuracy
Recency	24	104	23.07692
Recency, Number Agreement	37	104	35.57692
Recency, Number Agreement, Gender Agreement	81	104	77.88462
Recency, Number Agreement, Gender Agreement, Animistic Knowledge	89	104	85.57692

Table 4.5 – Results from experiment 5

Recency had a solid contribution of 23.5%. This is because the main character of the story Mr. Coyote is frequently mentioned by his full name and is almost always immediately followed by a pronoun referring back to him. For example: “Mr. Coyote took his…”

Number agreement contributes around 13% to the overall accuracy. Most of the resolutions that NA corrected over recency occurred near the end of the story. During this time “eyes” and “petals” are frequently being referred to, and NA helps sort out when those are being referred to since the majority of the pronouns are personal and possessive demonstrative pronouns referring to Mr. Coyote or Mr. Snail.

Gender agreement has a large contribution of 42%. This matches the trend observed in experiment one and experiment three that proposes that narrative styles of text with a main actor benefit greatly from the GA constraint source. In fact GA success in this story is focused primarily in the first three quarters where Mr. Coyote is the primary named actor in the text. Once Mr. Snail is introduced and referred to as frequently as Mr. Coyote, the performance of GA drops off. This drop can be described as GA losing its role as both the GA constraint source and the animistic knowledge constraint source since, as mentioned before, GA serves just as well as the animistic knowledge constraint when there is only one or very few gender identified main actors in a text.

Animistic knowledge has a small contribution of 8%. It is important to keep in mind that for most of the story animistic knowledge isn’t required because its cases are handled by gender agreement. The contribution of animistic knowledge picks up in the last quarter of the story and accounts for most of its contribution.

Two of the errors were failures to propose the correct antecedent for the demonstrative pronoun “This”. In this text both cases of “This” have an event as their correct coreferent, and since the recency proposal module only extracts nouns and noun phrases the pronoun resolution system never has a chance to correctly assign either “This” its correct coreferent. The solution is to improve the recency proposal module to include events as well as nouns and noun phrases.

Another two errors were failures to correctly resolve the coreferent of “it”. In both cases the pronoun resolution system is unable to extract enough information to identify the correct answer. This is because both instances of “it” are preceded by several candidates which are all compatible with “it” in terms of number agreement, don’t have any gender information to take advantage of, and which all pass the animistic knowledge constraint. Even addition of semantic knowledge wouldn’t solve these problems since both the correct and incorrect candidates belong to similar categories of objects (i.e. they are both things that have a definite size), so there isn’t enough context in the surrounding text to decide whether or not they make sense as possible coreferents. The probable solution is to include a module in the pronoun resolution system that keeps track of what object is the likely focus of discussion. This way the candidates that are not being discussed but are still being proposed could have their confidence scores lowered to the point where the correct coreferent is chosen.

The final ten errors occurred when the text frequently alternated referring to Mr. Snail and Mr. Coyote. In these cases gender agreement was unable to remove one or the other from the list of candidates because they are both male actors. They both always satisfy the number agreement filter for the same pronouns because they are both singular. Finally, they both always pass the animistic knowledge filter for the same pronouns because they are both animate beings. The result is the pronoun resolution system is reduced to using only the recency module to decide which actor is the correct corefernt, and because they are mentioned in an alternating pattern it is frequently incorrect.

These data suggest that the previous observation that narrative styles of text can be handled by very few constraint sources is correct. Although the sentence complexity is lower than the text in experiment three, the number of actors and references to them are more complex. Gender agreement is still the highest contributing module even though its performance drops off near the end of the text. The behavior of the GA module’s performance once again illustrates how the fewer actors there are in a story, the more it overlaps with animistic knowledge’s domain.

5. Conclusion and Directions for Future Work -

It is currently unclear exactly what types of information used as constraint sources are responsible for the success of other pronoun resolution systems. Further, it is relatively untested whether other systems succeed because they are well designed for the text they operate on or because they are a general solution to the problem of pronoun resolution. We have conducted experiments to determine the contribution of different constraint sources to pronoun resolution on different styles of written text. With this data we have proposed an experimental framework that sets up future research into how the writing style of a text document impacts the effectiveness of different constraint sources. The framework suggests that texts can be organized into categories such as sentence structure complexity, narrative, informative, anomalous, etc. Pronoun resolution accuracy in some styles of text was observed to benefit greatly from a subset of the available constraint sources. When our system did fail, it was unclear exactly what was required to improve performance further. Many times augmented semantic knowledge would have solved the problem. However it is also possible that more sophisticated syntactic analysis would also have sufficed. For example, a parallel structure module that takes advantage of the knowledge that a pronoun acting as the subject of a sentence usually has a subject of another sentence as a coreferent, a pronoun acting as a direct object usually refers to a direct object, etc. could increase the accuracy of the system but would be relatively easy to implement. This parallel structure module would also have very good general applicability to a wide domain. However if the initial syntactic modules added fail to increase the accuracy of the pronoun resolution system it is likely that augmenting the semantic knowledge would be much easier than trying to design the sophisticated heuristics of advanced syntactic methods. It's also likely that the augmented semantic approach would have a larger domain than the specialized syntactic approach because, for most applications, some sort of semantic analysis is going to be required anyway. This means that a semantic approach will have a larger set of text to which it can be effectively applied than a very specific syntactic heuristic designed to excel on a single style of text.

The next step is to pair more constraint sources with the writing styles for which they contribute the most to the accuracy of the pronoun resolution system. It would also be useful to design heuristics for certain styles of sentence structure. After that it will be necessary to design a method for examining a text and assigning it a to a writing style category. This process can be started by adapting preexisting methods for automatically deciding the sentence complexity, reading difficulty, and number of significant actors to be kept track of. This way the program can be designed to dynamically alter the importance it assigns the information provided by different constraint sources based on their observed effectiveness when operating on the style of text being analyzed. Finally devoting efforts to reducing the amount of information required from preprocessing will increase the robustness of the program by allowing to operate on text were syntactic information is sparse.

References

Anna Kupsc, Teruko Mitamura, Benjamin Van Durme, Eric Nyberg, "Pronominal Anaphora Resolution for Unrestricted Text", *International Conference on Language Resources and Evaluation* (LREC 2004) http://www.cs.cmu.edu/~ehn/JAVELIN/papers/javelin/671-anaphora.pdf

Shalom Lappin, Herbert J. Leass, “An Algorithm for Pronominal Anaphora Resolution” *Computational Linguistics* (Vol. 20, Num. 4, pp. 535-561, 1994) http://citeseer.ist.psu.edu/lappin94algorithm.html

C. Kennedy, B. Boguraev, “Anaphora for everyone: Pronominal anaphora resolution without a parser”, *The 16th International Conference on Computational Linguistics* (1996a) http://citeseer.ist.psu.edu/390979.html

R. Watson, J. Preiss, E.J. Briscoe, “The Contribution of Domain-independent Robust Pronominal Anaphora Resolution to Open-Domain Question-Answering”, *Proc of Int. Symposium on Reference Resolution and its Application to Question-Answering and Summarisation* (2003) http://www.cl.cam.ac.uk/~ejb/papers.html

J. L. Vicedo, A. Ferrández, “Importance of Pronominal Anaphora resolution in Question Answering systems”, *The 38th Annual Meeting of the Association for Computational Linguistics* (pp. 555-562, 2000) http://portal.acm.org/citation.cfm?id=1075288&dl=GUIDE&coll=GUIDE

Elaine Rich, Susann LuperFoy, “An Architecture for Anaphora Resolution”, *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing* (pp. 18-24, 1988) http://citeseer.ist.psu.edu/rich88architecture.html

Michael Schirber, “Dancing Bees Speak in Code”, *LiveScience Animal World* (27 May, 2005) http://www.livescience.com/animalworld/050527_bee_dance.html

Carol Moore, “Pirate’s Treasure”, *Children’s Storybooks Online* (2005) http://www.magickeys.com/books/pirate/index.html

Storie-Jean Agapith, “Mr. Coyote Meets Mr. Snail”, *Children’s Storybooks Online* (2005) http://www.magickeys.com/books/coyote/index.html

Christina Dodd, Lost in Your Arms, (pp. 123-124) Published by: Avon Books, 10 East 53rd Street, New York, New York 10022, 2002

Jane Austen, Persuasion, *The Complete Novels of Jane Austen* (pp. 1254-1255) Published by: Penguin Books Inc., 182-190 Wairau Road, Auckland 10, New Zealand, 1983